# Clinical Research Informatics Contributions from 2015

C. Daniel[1, 2], R. Choquet[1, 3], Section Editors for the IMIA Yearbook Section on Clinical Research Informatics
[1] INSERM UMRS 1142, Paris, France
[2] Direction of Information Systems, AP-HP, Paris, France
[3] BNDMR, Necker Hospital for Children, AP-HP, Paris, France

## Summary

**Objectives**: To summarize key contributions to current research in the field of Clinical Research Informatics (CRI) and to select best papers published in 2015.

**Method**: A bibliographic search using a combination of MeSH and free terms search over PubMed on Clinical Research Informatics (CRI) was performed followed by a double-blind review in order to select a list of candidate best papers to be then peer-reviewed by external reviewers. A consensus meeting between the two section editors and the editorial team was finally organized to conclude on the selection of best papers.

**Results**: Among the 579 returned papers published in the past year in the various areas of Clinical Research Informatics (CRI) - i) methods supporting clinical research, ii) data sharing and interoperability, iii) re-use of healthcare data for research, iv) patient recruitment and engagement, v) data privacy, security and regulatory issues and vi) policy and perspectives - the full review process selected four best papers. The first selected paper evaluates the capability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM) to support the representation of case report forms (in both the design stage and with patient level data) during a complete clinical study lifecycle. The second selected paper describes a prototype for secondary use of electronic health records data captured in non-standardized text. The third selected paper presents a privacy preserving electronic health record linkage tool and the last selected paper describes how big data use in US relies on access to health information governed by varying and often misunderstood legal requirements and ethical considerations.

**Conclusions**: A major trend in the 2015 publications is the analysis of observational, "nonexperimental" information and the potential biases and confounding factors hidden in the data that will have to be carefully taken into account to validate new predictive models. In addition, researchers have to understand complicated and sometimes contradictory legal requirements and to consider ethical obligations in order to balance privacy and promoting discovery.

## Keywords

Medical informatics, clinical research informatics, biomedical research, patient selection, phenotyping

Yearb Med Inform 2016:219-23
http://dx.doi.org/10.15265/IY-2016-044
Published online November 10, 2016

## Introduction

As pointed out by the survey of the Clinical Research Informatics (CRI) section, a major challenge identified in the 2015 publications in the field, is the growth of the diversity and size of data resources and the current trend to expand the underlying core technologies to enable big data analytics and data mining. New predictive models using large number of data may point out unobserved factors, unexpected findings or outcomes - this year's theme for the Yearbook – and highlight multifaceted aspects, which can be of great interest in health care. However, since big data enables the analysis of observational, "nonexperimental" information, the potential biases and confounding factors hidden in the data will have to be carefully taken into account to validate new predictive models [2]. The goal of this section is to provide an overview of relevant research published in the past year in the various areas of Clinical Research Informatics (CRI): i) methods supporting clinical research, ii) data sharing and interoperability, iii) re-use of healthcare data for research, iv) patient recruitment and engagement, v) data privacy, security and regulatory issues and vi) policy and perspectives.

## About the Paper Selection

A comprehensive review of published articles in 2015 addressing a wide range of issues for clinical research informatics was conducted. The selection was performed by querying Pubmed/Medline (from NCBI, National Center for Biotechnology Information) with a set of predefined keywords: Biomedical Research, Clinical research, Medical research, Pharmacovigilance, Patient Selection, Phenotyping, Genotype-phenotype associations, Data Collection, Epidemiologic Research Design, Epidemiologic Study Characteristics as Topic, Epidemiological Monitoring, Evaluation Studies as Topic, Clinical Trials as Topic, Feasibility Studies. References addressing topics of other sections of the Yearbook, such as Translational Bioinformatics were excluded based on predefined exclusion keywords such as Genetic Research, Gene Ontology, Human Genome Project, Stem Cell Research or Molecular Epidemiology.

Databases were searched on January 10, 2016 for papers published in 2015, considering the electronic publication date. From an original set of 579 references, a first subset of 406 references was considered according to the relevancy to the CRI field and blindly reviewed by the two section editors based on title and abstract. The articles were classified into several CRI categories and their contribution to CRI was rated as low, medium or high. Then, the two lists of references were merged, yielding 191 references that were classified as high contribution to CRI by at least one reviewer or medium contribution by both reviewers. The 191 references were reviewed by the two section editors jointly to select a

consensual list of 16 candidate best papers representative of all CRI categories. Following the IMIA Yearbook process, these 16 papers were peer-reviewed by editors and external reviewers (at least four reviewers per paper). Four papers were finally selected as best papers (Table 1). A content summary of these selected papers can be found in the appendix of this synopsis.

## Conclusion and Outlook

***Methods supporting clinical research:*** Big data is data "whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it". From an IT point of view, Bellazzi et al, reviewed the main types of big-data oriented solutions that are now available and increasingly adopted - cloud computing, parallel programming and new database technologies - and describe the research efforts carried on in the MOSAIC project in diabetes care, funded by the European Commission [2]. New probabilistic models for large-scale discovery of computational models of disease, or phenotypes can be developed [14]. In the domain of pharmacovigilance, Koutkias et al propose a semantically enriched framework to facilitate seamless access and use of different data sources and computational methods in an integrated fashion, bringing a new perspective for large-scale, knowledge-intensive signal detection [9].

***Data sharing and interoperability:*** Data heterogeneity is one of the critical problems in analyzing, reusing, sharing or linking datasets. For clinical trials, an electronic and structured study representation format used throughout the whole study life span can improve communication and potentially lower total study costs. The **first selected paper** from Huser et al [6], demonstrates the capability of the current version of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM) (1.3.2) to support the representation of case report forms (in both the design stage and with patient level data) during a complete clinical study lifecycle in the Intramural Research Program of the National Institutes of Health [6]. Several studies relate standard-based solutions providing a uniform access endpoint to patient structured and unstructured data. An approach exploiting well-established standards (such as HL7 v3), medical vocabularies (such as SNOMED CT, LOINC and HGNC) and Semantic Web technologies has been successfully tested to enable semantic interoperability in multi-centric clinical trials on breast cancer within the INTEGRATE and EURECA EU projects [1].

Standardizing data structure (through the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)), content (through a standard vocabulary with source code mappings), and analytics can enable an institution to apply a network-based approach to observational research across multiple, disparate observational health databases [16]. An alternative approach for data integration consists in semi-automated transformation of heterogeneous data sources to Web Ontology Language (OWL) formatted files by using a pre-defined model to extract the metadata. This approach was tested to link the South London Stroke Register (SLSR), the London Air Pollution toolkit (LAP) and the Clinical Practice Research Datalink (CPRD). The authors demonstrated that data can be faithfully converted to more suitable formats for further analysis [11]. The French national minimum data set (F-MDS-RD), composed of 58 data elements based on HL7 standard and various ontologies for diagnosis or sign encoding, was defined through national consensus and aligned with other similar initiatives for rare diseases, thus facilitating potential interconnections between rare disease registries [3]. An interesting approach and tool – BiobankConnect – was developed to significantly speed up the biobank harmonization process using ontological and lexical indexing [11]. Semi-automatically matches between 32 desired data elements of EU-BioSHaRE biobanks and ontology terms from BioPortal were computed and human curated.

***Re-use of healthcare data for research:*** **the second selected paper** from Kreuzthaler et al. describes a prototype for secondary use of electronic health records (EHR) data captured by non-standardized text, which is an important part of information representation in clinical information systems. Their prototypical information extraction system achieved an F-measure of 0.91 (precision=0.90, recall=0.93) for the training set and an F-measure of 0.90 (precision=0.89, recall=0.92) for the test set [10]. Focusing on EHR structured data, Mate et al. present an ontology-supported approach for data integration between clinical and research systems. Ontologies are used to organize and describe the medical concepts of both source and target system. Declarative transformation rules within ontologies are defined and used to automatically generate SQL code to extract data from source systems [12]. Biorepositories linked to de-identified EHR data have the potential to comple-

**Table 1** Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2016 in the section 'Clinical Research Informatics'. The articles are listed in alphabetical order of the first author's surname.

| Section<br>Clinical Research Informatics |
|---|
| ▪ Gray EA, Thorpe JH. Comparative effectiveness research and big data: balancing potential with legal and ethical considerations. J Comp Eff Res 2015 Jan;4(1):61-74. |
| ▪ Huser V, Sastry C, Breymaier M, Idriss A, Cimino JJ. Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM). J Biomed Inform 2015 Oct;57:88-9. |
| ▪ Kho AN, Cashy JP, Jackson KL, Pah AR, Goel S, Boehnke J, Humphries JE, Kominers SD, Hota BN, Sims SA, Malin BA, French DD, Walunas TL, Meltzer DO, Kaleba EO, Jones RC, Galanter WL. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. J Am Med Inform Assoc 2015 Sep;22(5):1072-80. |
| ▪ Kreuzthaler M, Schulz S, Berghold A. Secondary use of electronic health records for building cohort studies through top-down information extraction. J Biomed Inform 2015 Feb;53:188-95. |

ment traditional epidemiologic studies in genotype-phenotype studies of complex human diseases and traits. Unlike traditional epidemiologic data, EHR data are highly variable across patients and often available within unstructured clinical notes and therefore, the development of algorithms to extract phenotypes for analysis is needed. Dumitrescu et al. explored the impact that algorithm decision logic has on genetic association study results for a single quantitative trait, high-density lipoprotein cholesterol (HDL-C). The study suggests that, at least for this quantitative trait, algorithm decision logic and phenotyping details do not appreciably affect genetic association study test statistics [4].

***Patient recruitment and engagement:*** Soto-Rey et al. assessed the value of obtaining potential study participant counts using an automated patient count cohort system for large multi-country and multi-site trials: the Electronic Health Records for Clinical Research (EHR4CR) system [15].

***Data privacy, security and regulatory issues:*** In the **third selected paper**, Kho et al. [8] proposed a privacy preserving electronic health record linkage tool used to successfully link and de-duplicate 7 million records from 6 hospitals in the Chicago area resulting in a cohort of 5 million unique records. They reduced duplication of patient records across sites by as much as 28%. They propose an innovative methodology to generate 17 different hashs from local EHRs in order to maximize exact matches as well as improving record linkage when some data is missing (such as the social security number). The proposed methodology achieved a sensivity of 96% and a specificity of 100%. This approach is of great interest to setup and evaluate public health surveillance from EHR data as opposed to classical population based registries. Kaye et al. [7] propose an innovative approach for involving participants in clinical research: the dynamic consent. The dynamic consent consists in a personalized, digital communication interface that connects researchers and participants and facilitates two-way communication to stimulate a more engaged, informed and scientifically literate participant population where individuals can tailor and manage their own consent preferences. This approach has mainly been developed in bio banking contexts but has potential application in other domains for a variety of purposes. **The last selected paper**, by Gray et al. [5] describes how big data use in US relies on access to health information governed by varying and often misunderstood legal requirements and ethical considerations related to patient privacy. This review is relevant, beyond the US, in the context of comparative effectiveness research conducted internationally at scale. In such context, researchers have to understand complicated and sometimes contradictory legal requirements and ethical consideration in order to properly balance privacy and discovery. Ensuring that patients fully understand how their data will be used and by whom, and more generally speaking consumer engagement, is a key factor for overcoming legal and ethical barriers to effective and robust use of big data in comparative research, clinical decision support and quality improvement.

**Acknowledgement**

We would like to acknowledge the support of Martina Hutter and the reviewers in the selection process of the IMIA Yearbook.

## References

1. Alonso-Calvo R, Perez-Rey D, Paraiso-Medina S, Claerhout B, Hennebert P, Bucur A. Enabling semantic interoperability in multi-centric clinical trials on breast cancer. Comput Methods Programs Biomed 2015 Mar;118(3):322-9.
2. Bellazzi R, Dagliati A, Sacchi L, Segagni D. Big Data Technologies: New Opportunities for Diabetes Management. J Diabetes Sci Technol 2015 Apr 24;9(5):1119-25
3. Choquet R, Maaroufi M, de Carrara A, Messiaen C, Luigi E, Landais P. A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research. J Am Med Inform Assoc 2015 Jan;22(1):76-85.
4. Dumitrescu L, Goodloe R, Bradford Y, Farber-Eger E, Boston J, Crawford DC. The effects of electronic medical record phenotyping details on genetic association studies: HDL-C as a case study. BioData Min 2015 May 6;8:15.
5. Gray EA, Thorpe JH. Comparative effectiveness research and big data: balancing potential with legal and ethical considerations. J Comp Eff Res 2015 Jan;4(1):61-74.
6. Huser V, Sastry C, Breymaier M, Idriss A, Cimino JJ. Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM). J Biomed Inform 2015 Jul 15.
7. Kaye J, Whitley EA, Lund D, Morrison M, Teare H, Melham K. Dynamic consent: a patient interface for twenty-first century research networks. Eur J Hum Genet 2015 Feb;23(2):141-6.
8. Kho AN, Cashy JP, Jackson KL, Pah AR, Goel S, Boehnke J, et al. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. J Am Med Inform Assoc 2015 Sep;22(5):1072-80.
9. Koutkias VG, Jaulent MC. Computational approaches for pharmacovigilance signal detection: toward integrated and semantically-enriched frameworks. Drug Saf 2015 Mar;38(3):219-32.
10. Kreuzthaler M, Schulz S, Berghold A. Secondary use of electronic health records for building cohort studies through top-down information extraction. J Biomed Inform 2015 Feb;53:188-95.
11. Liang SF, Taweel A, Miles S, Kovalchuk Y, Spiridou A, Barratt B, et al. Semi automated transformation to OWL formatted files as an approach to data integration. A feasibility study using environmental, disease register and primary care clinical data. Methods Inf Med 2015;54(1):32-40.
12. Mate S, Köpcke F, Toddenroth D, Martin M, Prokosch HU, Bürkle T, Ganslandt T. Ontology-based data integration between clinical and research systems. PLoS One 2015 Jan 14;10(1):e0116656.
13. Pang C, Hendriksen D, Dijkstra M, van der Velde KJ, Kuiper J, Hillege HL, et al. BiobankConnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing. J Am Med Inform Assoc 2015 Jan;22(1):65-75..
14. Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, Elhadad N. Learning probabilistic phenotypes from heterogeneous EHR data. J Biomed Inform 2015 Dec;58:156-65.
15. Soto-Rey I, Trinczek B, Girardeau Y, Zapletal E, Ammour N, Doods J, et al. Efficiency and effectiveness evaluation of an automated multi-country patient count cohort system. BMC Med Res Methodol 2015 May 1;15:44.
16. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc 2015 May;22(3):553-64.

Correspondence to:
Christel Daniel, MD, PhD
INSERM UMRS 1142 - WIND-DSI
– Assistance Publique – Hôpitaux de Paris
05 rue Santerre
75 012 Paris, France
Tel: +33 1 48 04 20 29
E-mail: christel.daniel@ aphp.fr

# Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2016, Section 'Clinical Research Informatics'

### Huser V, Sastry C, Breymaier M, Idriss A, Cimino JJ

### Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM)

### J Biomed Inform 2015;57:88-99

Efficient communication of a clinical study protocol and case report forms during all stages of a human clinical study is an important concern in the Clinical Research Informatics field. From an institutional perspective, the benefits of using an electronic and structured study representation format depend on the overall volume of studies, the proportion of studies eventually submitted to FDA and on the enterprise electronic research systems used at a given site. Standardization benefits are most apparent for multi-site studies using multiple non-centralized electronic data capture systems and for studies later submitting data to the FDA or to data sharing platforms where clinical research data are stored and re-analyzed by investigators. The most relevant standard for representing clinical study data – the Operational Data Model (ODM) of the Clinical Data Interchange Standards Consortium (CDISC) – was initially developed for the exchange of case report forms data but an ODM extension called Study Design Model, introduced in 2011, provides additional protocol representation elements. The authors evaluated ODM's ability to capture all necessary protocol elements during a complete clinical study lifecycle in the Intramural Research Program of the National Institutes of Health. They picked the most complex study and the objective was to demonstrate that the adoption of ODM to standardize the representation of protocol metadata elements did not alter their capacity to offer to the study principal investigator the possibility to efficiently use his/her study for an informatics analysis. For each study stage, the authors presented a list of limitations in the ODM standard and identified necessary vendor or institutional extensions. Additional needed ODM extensions would: i) better address the standardization of protocol-level metadata; ii) consider the requirements of observational, non-regulated studies (addition of more user roles, study types and phases); iii) support more restricted syntax for CRF data validation; iv) ease the annotation of CRF questions with coded concepts from any terminology (e.g., SNOMED or LOINC) or data element definition scheme (e.g., Common Data Elements) and v) promote the use of ODM by healthcare institutions and EHR vendors (through ODM mediators or implementation of the Retrieve Form for Data-capture (RFD) profile maintained by the Integrating Healthcare Enterprise (IHE) inititative).

### Kho AN, Cashy JP, Jackson KL, Pah AR, Goel S, Boehnke J, Humphries JE, Kominers SD, Hota BN, Sims SA, Malin BA, French DD, Walunas TL, Meltzer DO, Kaleba EO, Jones RC, Galanter WL

### Design and implementation of a privacy preserving electronic health record linkage tool in Chicago

### J Am Med Inform Assoc 2015 Sep;22(5):1072-80

Along with the widespread adoption of electronic health record solutions in US hospitals, a great volume of restrospective clinical data is available. Public health officials and researchers have raised an interest in getting efficient, yet preserving anonymity, ways to link hospital clinical records across health care institutions. Since no single and permanent patient identifier exists in the US, and very few implementations of master patient index service is used, a specific method and software is required. The authors present their real-world implementation of a Distributed Common Identity for the Integration of Regional Health Data (DCIFIRHD) across 6 healthcare institutions within the Chicago region. Each healthcare institution provided pre-exitsting data from local datawarehouses or medical record systems. Global IRB approval for the study to be conducted was set in 12 months. Datasets were composed of patients aged from 18 to 89 years that have visited the hospital between 2006 and 2012. Specific sensible diagnoses were excluded (HIV, etc.). The experimentation was done on structured data only due to potential re-identification risk linked to the use of hospital discharge records (free text). Data was pre-processed at local level first, through the DCIFIRHD application, an external trusted party was responsible for assigning a specific project passphrase and passcode to hash outgoing data for each hospital in order to protect against dictionary attacks. To ensure full HIPAA-compliance, after linkage, each cluster of site-specific PatientIDs was replaced with nonderived StudyID for use in subsequent analyses. Up to 17 512-bit hashes were generated for each patient from personal data combining first name, last name, date of birth and/or social security number (SSN) in different combinations. Three different methods were used to hash the combined variables: normal SHA 512 hash, first three letters of first and last names and hash, Soundex. The matching method used a simple deterministic algorithm. As a result, the reducted percentage of patient duplicates ranged from 10.9% to 28%. Age group comparison of the study and census data showed that the proportion of patients by 5-year bins was similar. The current implementation of the solution requires a central data collector hold by a central trusted provider, authors are looking to improve this through the PCROnet research initiative. The use of extra data could also improve the matching performance.

### Kreuzthaler M, Schulz S, Berghold A

### Secondary use of electronic health records for building cohort studies through top-down information extraction

### J Biomed Inform 2015 Feb;53:188-95

Bridging the gap between *patient–based storage systems* (clinical information systems) and *disease-related search systems* (biobanks) to build cohort studies is a key clinical research informatics challenge. Building cohort studies within a clinical setting requires to gather and integrate var-

ious forms of data including unstructured data. A cohort study is defined by authors as the investigation of the association between exposure and disease conducted by following individuals through a time span and measuring the rate of occurance of new cases in different exposition groups. In this paper, the authors first present a state of the art of text mining challenges within the healthcare application field as well as prototypes related to the presented work. The prototypical information extraction system developed by the authors processed heterogeneous (50% of free text) datasources within the hospital. Within the free text, the authors showed that most of information is expressed using a "unit/value" pair model. The Apache Unstructured Information Management Architecture was used together with *simple* regular expressions to process with information extraction. They obtained a F-measure of 0.91 for the training set and 0.90 for the test set using a gold standard. Despite the promising results in the field of secondary use of clinical data, the authors hightlight the need to increase the quality of clinical documentation within EHRs and to refine methods to capture the context of extracted data (body height measured by the patient or a nurse).

## Gray EA, Thorpe JH

## Comparative effectiveness research and big data: balancing potential with legal and ethical considerations

J Comp Eff Res 2015 Jan;4(1):61-74

Big data solutions have the potential to transform comparative effectiveness research, by facilitating the collection and aggregation of volumes of multi-source data to enable comparisons across care settings, patient populations, treatment combinations, payers and time. However, intensive secondary use of clinical data at large scale raises significant legal and ethical considerations governing privacy and security, which are complex and varied. Misconceptions and lack of understanding about the legal framework applicable to health information operate as a barrier to the full potential of big data. The authors give an overall view of the scope of some legal and ethical issues in US and how they may be managed to fully realize the potential of big data.

There is no overarching, comprehensive legal framework applicable to secondary use of health information across US. Federal and state laws and regulations governing health information often overlap and in some cases may contradict each other. Regarding to privacy and security, data minimization and data destruction after use are antithetical to big data best practices of collecting as much data as possible and holding on to it indefinitely. In addition, de-identified data is not useful for many purposes including determining causal relationships, conducting genetic research, provider performance measurement and contacting patients for various aspects of patient engagement. Moreover, the public does not trust that data can be truly anonymous and using de-identified data may be ethically questionable if it negatively affects patient-provider relationships. Regarding to patient consent, since there is no common consent architecture, different consent elements and processes may apply depending on the situation and patients may not fully understand how their data will be used and by whom. In addition seeking consent may be impracticable and can bias results.

Future perspectives are i) to develop consistent framework for patient consent tailored to specific patients and research and supported by clinical decision support systems and interactive online process; ii) to re-conceptualize de-identification and engage with federated databases that collect and share data according to specified research protocols in order to access more and better de-identified data; iii) to consider purpose-specific data minimization and iv) to educate patients on the value of research and ensure that security protocols are in place to protect data.

The authors consider that, as the technology is evolving, there is also a major shift in public perceptions of privacy that may fundamentally change the way society views confidentiality and the benefit of disclosure for the public good.